IBM Research

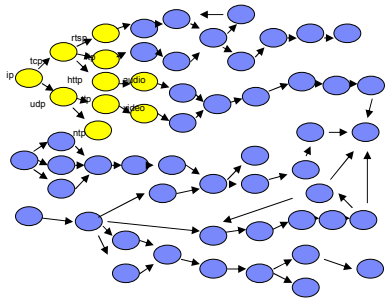
Large-Scale Video Semantic Filtering

Ching-Yung Lin

Exploratory Stream Processing Systems,
IBM T. J. Watson Research Center

November 16, 2005

Outline -- Large-Scale Video Semantic Filtering



- Introduction and Motivation
- Prior Arts on Video Semantic Classification
- Compressed-Domain Feature Extraction
- Complexity-Accuracy Trade-Off
- Speed Up SVM Classification
- Speech-Text Complexity-Accuracy Analysis
- Conclusions
- Prospective Projects

System Overview

Create breakthrough technology to enable

…large-scale production, analysis and management of

…information and knowledge from

…thousands of disparate input sources

…using an autonomic system that

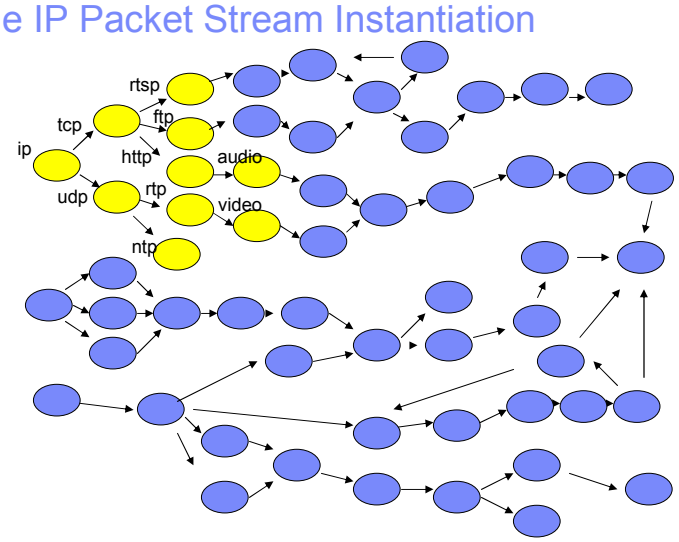
…continuously adapts processing to satisfy demands of

…changing environment, content, and requests.

What is Large-scale?

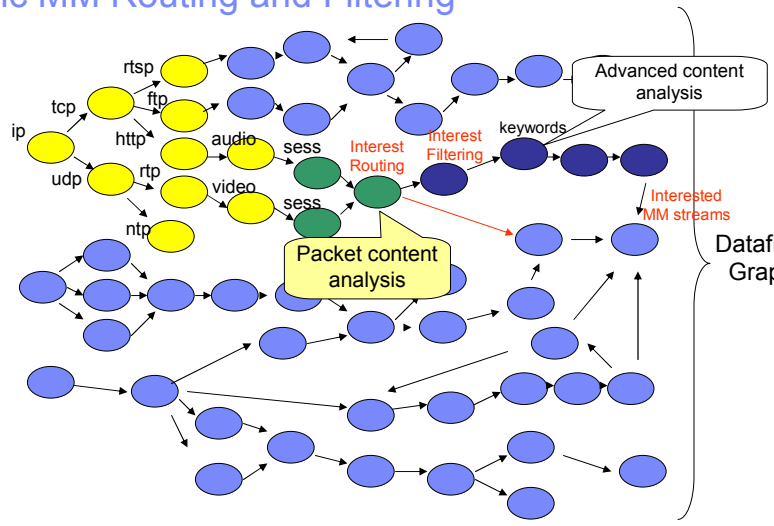
- 10Gbit/s Continuous Feed Coming into System
- Types of Data
  - Speech, text, moving images, still images, coded application data, machine-to-machine binary communication
- System Mechanisms
  - Telephony: 9.6Gbit/sec (including VoIP)
  - Internet
    - ✓ Email: 250Mbit/sec (about 500 pieces per second)
    - ✓ Dynamic web pages: 50Mbit/sec
    - ✓ Instant Messaging: 200Kbit/sec
    - ✓ Static web pages: 100Kbit/sec
    - ✓ Transactional data: TBD
  - TV: 40Mb/sec (equivalent to about 10 stations)
  - Radio: 2Mb/sec (equivalent to about 20 stations)

Example IP Packet Stream Instantiation



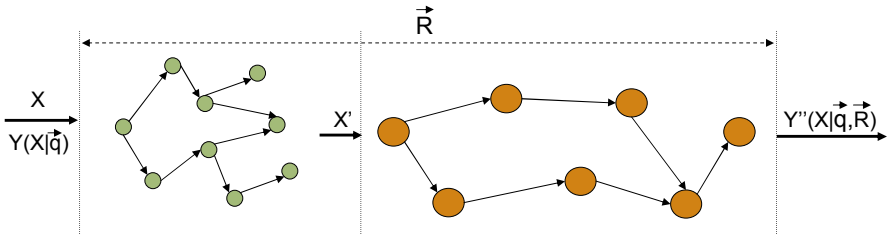
per PE rates	200-500MB/s	~100MB/s	10 MB/s

Semantic MM Routing and Filtering



per PE rates	200-500MB/s	~100MB/s	10 MB/s

Objective



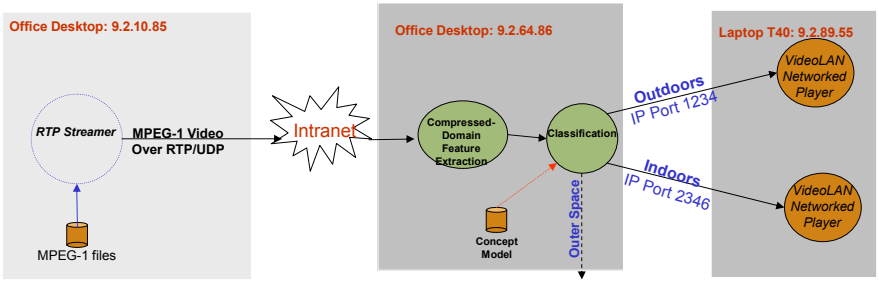
- **Input data X – Queries q – Resource R**
  - Y(X | q): Relevant information
  - Y'(X | q, R) ∈ Y(X | q): Achievable subset given R
- **Configurable Parameters of Processing Elements to maximize relevant information:**
  - Y''(X | q, R) > Y'(X | q, R),

  with resource constraint.
- **Required resource-efficient algorithms for:**
  - Classification, routing and filtering of signal-oriented data: (audio, video and, possibly, sensor data)

Video Concept Classification/Routing

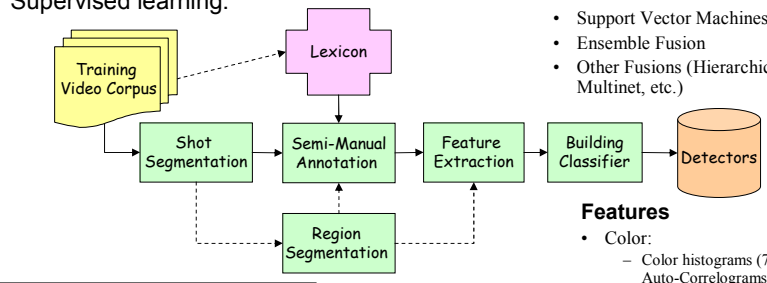
- **Objective:**
  - Build Moderate Amount of Real-Time Concept Classifiers, e.g., Human, Outdoors, Face, Indoors, etc.

- **Test-bed Corpus:**
  - NIST TRECVID 2003 corpus: 128 hours of MPEG-1 videos, 62 hours of them were manually annotated by 23 worldwide groups using IBM *VideoAnnEx* Collaborative Annotation System.
  - 38 hours of video for training; 24 hours (partitioned to 6-, 6-, 12-hour sets) for internal evaluation; 65 hours for NIST benchmarking.



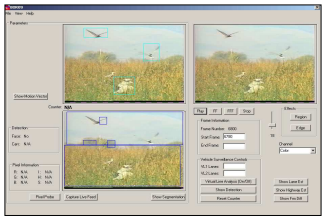
Prior Art: Supervised Learning for Building Generic Concept Detectors

Supervised learning:



Classification and Fusion:

- Support Vector Machines (SVM)
- Ensemble Fusion
- Other Fusions (Hierarchical, SVM, Multinet, etc.)

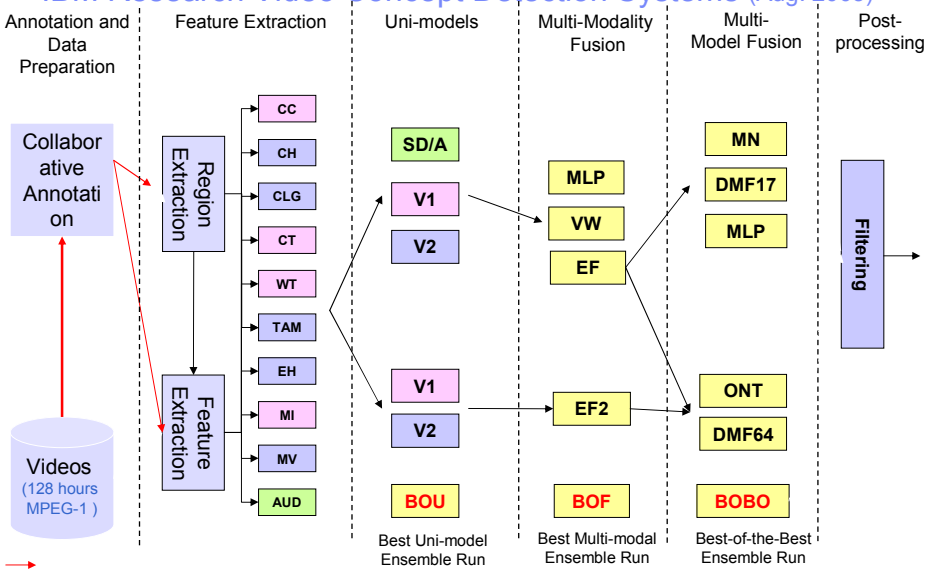


- **Regions**
  - Object (motion, Camera registration)
  - Background (5 lg regions / shot)

Features

- Color:
  - Color histograms (72 dim, 512 dim), Auto-Correlograms (72 dim)
- Structure & Shape:
  - Edge orientation histogram (32 dim), Dudani Moment Invariants (6 dim), Aspect ratio of bounding box (1dim)
- Texture:
  - Co-occurrence texture (48 dim), Coarseness (1 dim), Contrast (1 dim), Directionality (1 dim), Wavelet (12 dim)
- Motion:
  - Motion vector histogram (6 dim)

IBM Research Video Concept Detection Systems (Aug. 2003)



Prior Art: Classification Steps

MPEG-1 sequence

- detectors

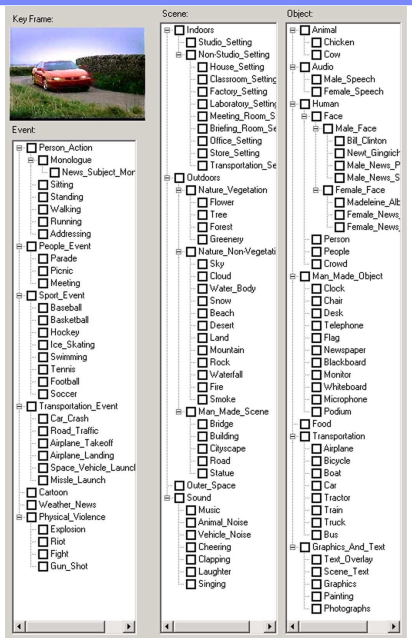
"Text-overlay"

"Outdoors"

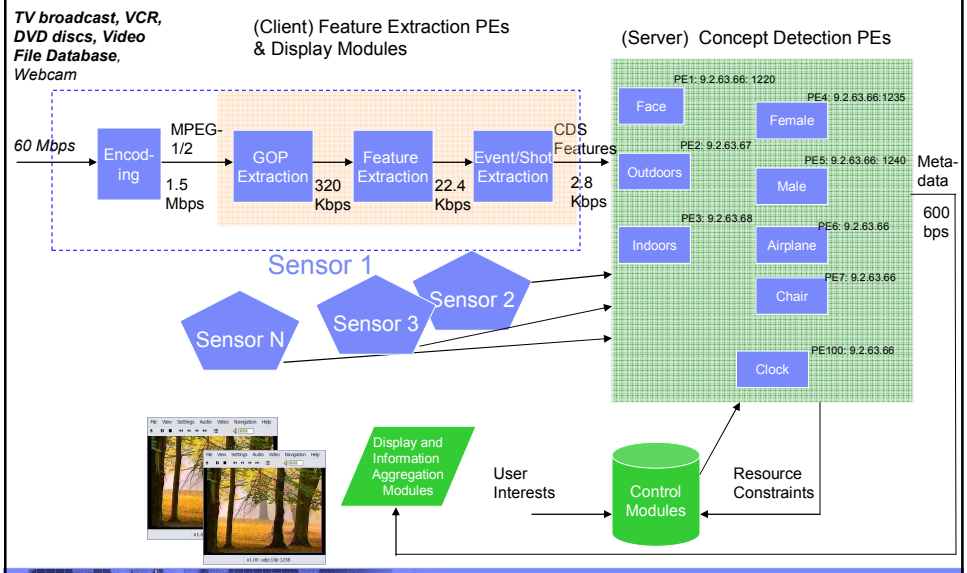
"People"

Novel Visual Semantic Concept Filters

- 100 visual filters were built (v.s. 64 in 2003).
- Concepts are arranged hierarchically.
- Based on novel compressed-domain sliced features.
- Neither segmentation nor fusion is used.



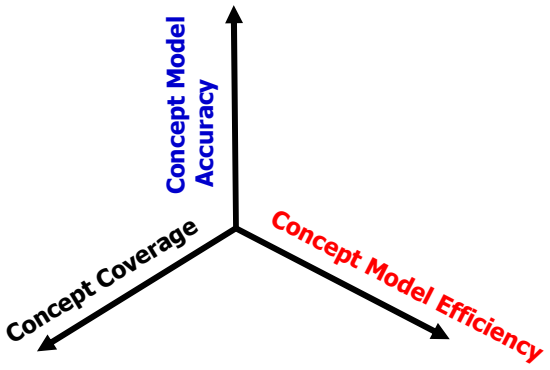
An Example of Video Sensor Network System



Main Process

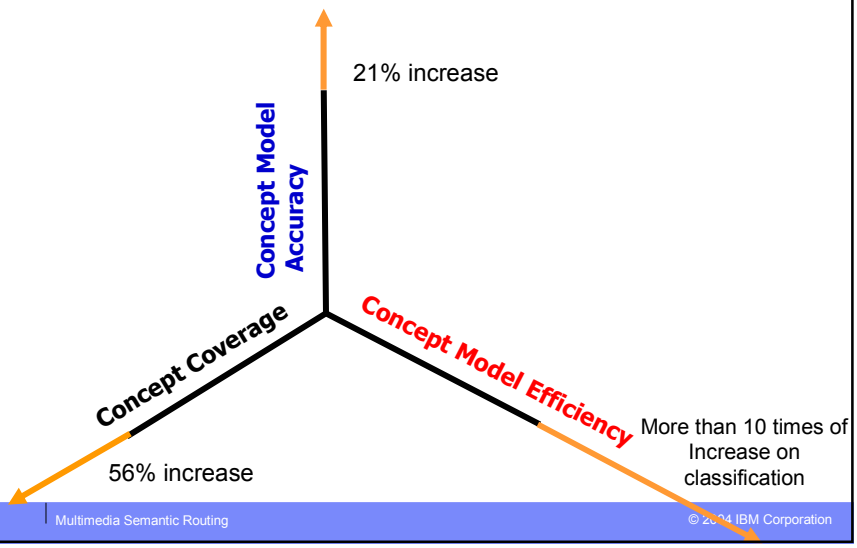
- This feature set is extracted as follows:
  1. Parse the MPEG-1/2 packets and get the beginning of an I-frame or the closest I-frame of a pre-specified shot keyframe.
  2. Using the VLC maps to map variable-length codes to the DCT-domain coefficients.
  3. Within an MPEG slice or a union of slices, truncate selected DCT coefficients and calculate the histogram of these DCT coefficients.
  4. Form a feature vector of the frame based on the histogram coefficients with multiple slices.
- In the above procedure, we can see that no multiplication operation is required to get these feature vectors. Only addition is needed for getting the histogram.
- In a typical situation, we partition a frame into three slices, and use 3 DCT coefficient histogram (1 DC and 2 lowest frequency AC coefficients) on all YCbCr domains. This forms a 576-dimensional feature vector.
- After feature extraction, SVM is used to train models and classification.

Three Measurements of Multimedia Semantic Concept Detectors

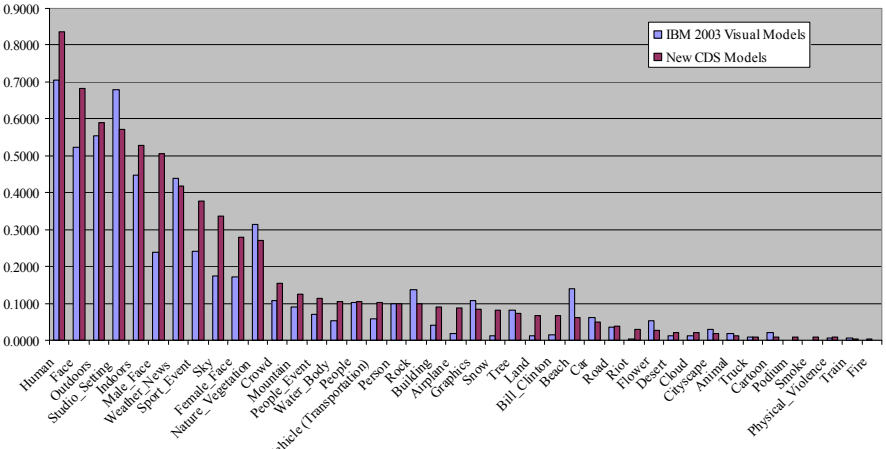


Progress on Novel Concept Filters for Distillery MM Semantic Routing

- **Baseline: (state-of-the-art) IBM TREC '03 Visual Detectors**



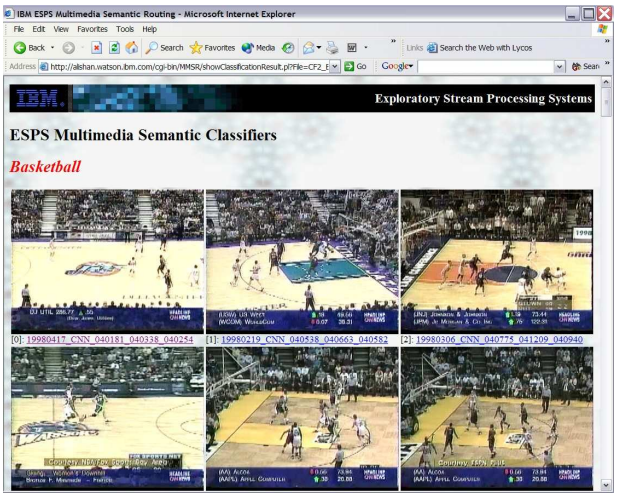
Performance Comparison of the Compressed-Domain Detectors and IBM 2003 Visual Detectors



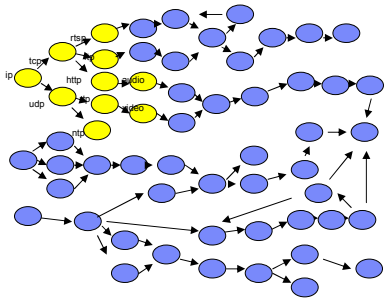
- Improved Mean Average Precision: 21.48%
- Improved Efficiency:
  - >2M multiplication operations needed for generating features for IBM-03 classification per key frame.
  - no multiplication operations needed for new CDS models, only 6K addition operations.

Demo -- Novel Semantic Concept Filters

- http://www.research.ibm.com/VideoDIG
- **E.g.:**



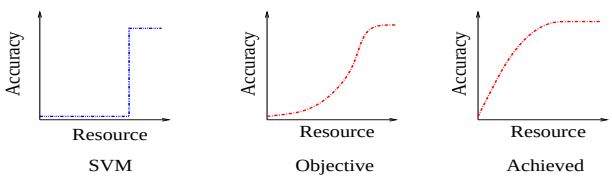
Outline -- Large-Scale Video Semantic Filtering



- Introduction and Motivation
- Prior Arts on Video Semantic Classification
- Compressed-Domain Feature Extraction
- Complexity-Accuracy Trade-Off
- Speed Up SVM Classification
- Speech-Text Complexity-Accuracy Analysis
- Conclusions
- Prospective Projects

Complexity Reduction Introduction

- **Objective: Real-time classification of instances using Support Vector Machines (SVMs)**
- **Computationally efficient and reasonably accurate solutions**
- **Techniques capable of adjusting tradeoff between accuracy and speed based on available computational resources**



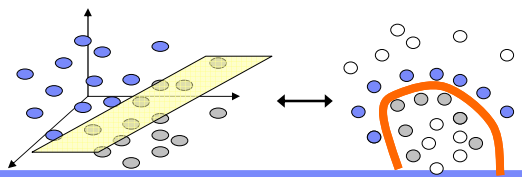
SVMs

- **Developed over the past decade**
- **Strong theoretical background**
  - Attempt to achieve Structural Risk Minimization
  - Minimize upper bound on generalization error instead of MSE
- **Impressive empirical results, handles noisy data**
- **Effective over diverse domains**
  - Text (character/word based), Image, Surveillance (Video/Network traffic), Bio (Gene/Protein/Disease) etc.

SVM formulation

- **Given :**
  - Training instances $\{\mathbf{x}_i\}$ with labels y_i
- **Objective :**
  - Find maximum margin hyperplane separating positive and negative training instances

SVM



Salient Features

- **Formulation:**

$$min_{\mathbf{w},b} \quad \frac{1}{2} \parallel \mathbf{w} \parallel^2$$

$$s.t. \quad y_i(\mathbf{w} \cdot \phi(x_i) + b) \geq 1$$

- **Optimization problem with quadratic objective function, linear constraints (QPP)**
- **Hyperplane separator in *projected* space**
- **Implicit projection of instances**

Kernel Projection

- **Implicit projection to feature space**

$$\mathrm{K}(\mathrm{x}_i, x_j) = < \phi(x_i), \phi(x_j) >$$

- **Measures similarity between instances in projected space**
- **Examples:**
  - Gaussian : $\mathrm{Exp}(\text{-}\gamma \parallel x_i - x_j \parallel^2)$
  - Laplacian : $\mathrm{Exp}(\text{-}\gamma \parallel x_i - x_j \parallel)$
  - Polynomial : $(1 + \mathrm{x}_i \cdot x_j)^p$

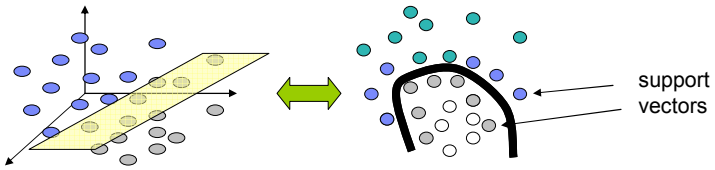
Decision

- **Score of unseen instance** u_j : $w \cdot \phi(u_j)$
- **In terms of Lagrangian multipliers**

$$\sum_i \alpha_i y_i k(x_i, u_j)$$

- **Computational Cost : O($n_{sv} d$)**
  - n_{sv}: Number of support vectors
  - d : Dimensionality of each data instance

Complexity Analysis on SVM classifiers



- **Support Vector Machine**
  - Largest margin hyperplane in the projected feature space
  - With good kernel choices, all operations can be done in low-dimensional input feature space

- SVM Classifier:

$$f(x) = \sum_{i=1}^{S} a_i \cdot k(x, x_i) + b$$

where *S* is the number of support vectors, k(.,.) is a kernel function. E.g., $k(x, x_i) = e^{-\frac{\|x - x_i\|}{r}}$

- Complexity *c*: operation (multiplication, addition) required for classification

$$c \propto S \cdot D$$

where *D* is the dimensionality of the feature vector

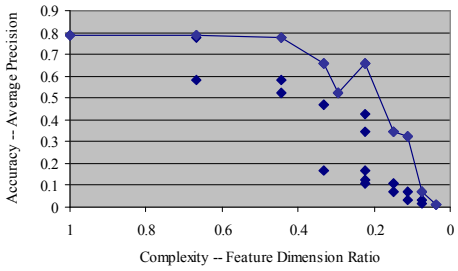
Problems

- **Number of support vectors grows quasi-linearly with size of training set [Tipping 2000]**
- **Inner product with each support vector of dimensionality d expensive**
  - Example TREC2003
    - Human : 19745 support vectors
    - Face : 18090
- **High data rates(10Gbits/sec) means large number of abandoned data**

Example

- **Processing Power 1 Ghz**
- **10000 support vectors**
- **1000 / 2 features per instance**
- **Order of at least 10^7 operations required per stream per sec**
- **Translates to less than 100 instances evaluated per sec with only one classifier**

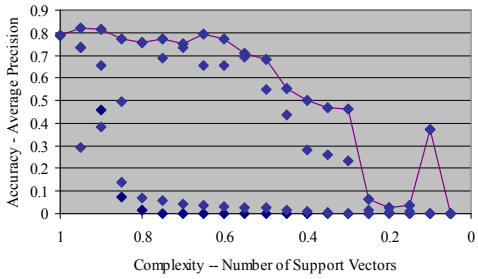
Naïve Approach I – Feature Dimension Reduction



- Experimental Results for Weather_News Detector
- Model Selection based on the Model Validation Set
- E.g., for Feature Dimension Ratio 0.22, (the best selection of features are: 3 slices, 1 color, 2 texture selections), the accuracy is decreased by 17%.

Slice	Color	Texture	Feature Dimension Ratio	AP
3	3	3	1	0.7861
3	3	2	0.666666667	0.7861
3	2	3	0.666666667	0.7757
2	3	3	0.666666667	0.5822
3	2	2	0.444444444	0.7757
2	3	2	0.444444444	0.5822
2	2	3	0.444444444	0.5235
3	3	1	0.333333333	0.4685
3	1	3	0.333333333	0.6581
1	3	3	0.333333333	0.1684
2	2	2	0.296296296	0.5235
3	2	1	0.222222222	0.427
3	1	2	0.222222222	0.6581
2	3	1	0.222222222	0.1241
2	1	3	0.222222222	0.3457
1	3	2	0.222222222	0.1684
1	2	3	0.222222222	0.1065
2	2	1	0.148148148	0.0699
2	1	2	0.148148148	0.3457
1	2	2	0.148148148	0.1065
3	1	1	0.111111111	0.3219
1	3	1	0.111111111	0.0314
1	1	3	0.111111111	0.07
2	1	1	0.074074074	0.0318
1	2	1	0.074074074	0.0173
1	1	2	0.074074074	0.07
1	1	1	0.037037037	0.0123

Naïve Approach II – Reduction on the Number of Support Vectors



- Proposed Novel Reduction Methods:
  - **Ranked Weighting**
  - **P/N Cost Reduction**
  - **Random Selection**
  - **Support Vector Clustering and Centralization**
- Experimental Results on Weather_News Detectors show that complexity can be at 50% for the cost of 14% decrease on accuracy

Approach Overview

- **Reduced set methods**
- **Focus on support vectors and weights already obtained (Same SVM formulation)**
- **Accommodate effect due to all support vectors in approximation**
- **Amenable to**
  - multiple kernels
  - L_1, L_2 norms
- **Approximate support vector scoring with consideration for overall error**

Approaches

- **Weighted Clustering Approach:**
  - Approximate effect of support vector(s) by (possibly hypothetical) instance with weight adjustment
- **Error Radius Approach:**
  - Determine *error radius* and approximate effect outside same

Weighted Clustering Approach

- **Basic steps**
  - Cluster support vectors
  - Use cluster center as representative for all support vectors in cluster
  - Determine scalar weight associated with each cluster center
  - Use only cluster centers to score new instances

Weighted Clustering Approach (contd.)

- **Clustering**
  - Cluster positive and negative support vectors separately
  - Ratio of +ve clusters to -ve clusters same as ratio of +ve support vectors to -ve support vectors
  - Attempts to avoid *bloated* clusters because of imbalance
  - Manual decision on number of clusters

Weighted Clustering Approach (contd.)

- **Cluster center weight**
  - Given :
    - Support vector $\mathbf{x}_i$ in cluster with cluster center $\mathbf{x}_c$
    - Weight α_i associated with $\mathbf{x}_i$
  - Δ_i Squared distance between $\mathbf{x}_i$ and $\mathbf{x}_c$
  - Unknown instance at squared distance
    - d from $\mathbf{x}_c$
    - $d - \delta_i$ from $\mathbf{x}_i$

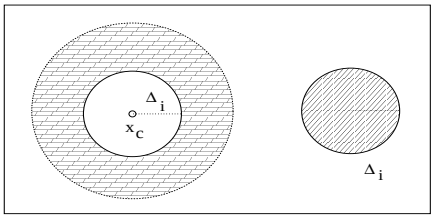
Cluster center weight (contd.)

- **Difference in scores**
  $\gamma_i \; exp(-\frac{d}{p}) - \alpha_i exp(-\frac{(d-\delta_i)}{p})$
- **Optimal** $\gamma_i = \alpha_i exp(\frac{\delta_i}{p})$
- **Problem:**
  - Determined γ_i specific to instance
  - Impossible to obtain speedup if γ_i computed on a per-instance basis
- **Solution:**
  - Find γ_i minimizing error on average

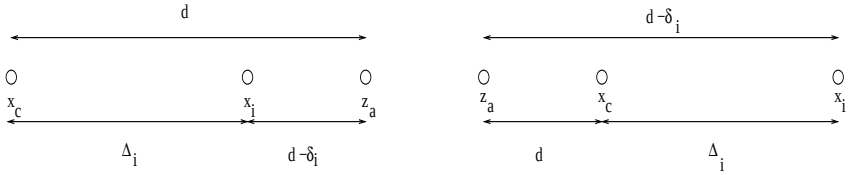
Cluster center weight (contd.)

- **Choose γ_i minimizing square of difference in scores over all δ_i and *d***
- **Sub-cases :**

$d \geq \Delta_i$ $\qquad$ $d < \Delta_i$



Cluster center weight (contd.)

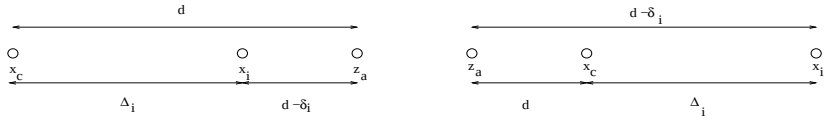


- **Case 1: Distance to unseen instance** $d \geq \Delta_i$

$$\int_{\Delta_i}^{\infty} \int_{-\Delta_i}^{\Delta_i} (\gamma_i exp(-\frac{d}{p}) - \alpha_i exp(-\frac{(d-\delta_i)}{p}))^2 \mathbf{d}\delta_i \mathbf{d}d.$$

- **Optimal** $\gamma_{i_{upper}} = \alpha_i p \frac{exp(\frac{\Delta_i}{p}) - exp(-\frac{\Delta_i}{p})}{2\Delta_i}$
- **Minima verification:** $2\Delta_i p exp(-\frac{2\Delta_i}{p}) \geq 0$
- **Optimal $\gamma_{i_{upper}}$ obtained for** $d \geq \Delta_i$

Cluster center weight (contd.)



- **Case 2:** $d < \Delta_i$

$$\int_0^{\Delta_i} \int_{\Delta_i - 2d}^{\Delta_i} (\gamma_i exp(-\frac{d}{p}) - \alpha_i exp(-\frac{(d+\delta_i)}{p}))^2 \mathbf{d}\delta_i \mathbf{d}d$$

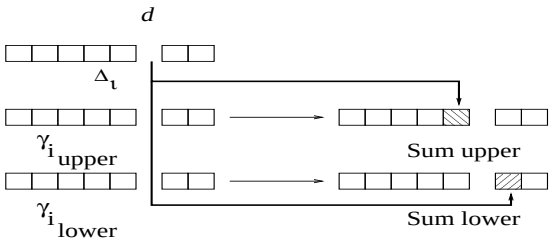
- **Optimal** $\gamma_{i_{lower}} = \frac{\alpha_i exp(-\frac{\Delta_i}{p})[\frac{2\Delta_i}{p} - 1 + exp(-\frac{2\Delta_i}{p})]}{1 - exp(-\frac{2\Delta_i}{p}) - \frac{2\Delta_i}{p} exp(-\frac{2\Delta_i}{p})}$
- **Minima :** $exp(\frac{2\Delta_i}{p}) > (1 + \frac{2\Delta_i}{p})$
- **Optimal** $\gamma_{i_{lower}}$ **obtained for** $d < \Delta_i$

Using the weights

- **For every support vector in cluster**
  - Distance Δ_i known
  - Two weights computed
- **Cumulative effect of all support vectors in clusters additive**
  - Δ_i because of various support vectors added up at center to simulate effect of all support vectors
- Δ_i **sorted, weight arrays rearranged**

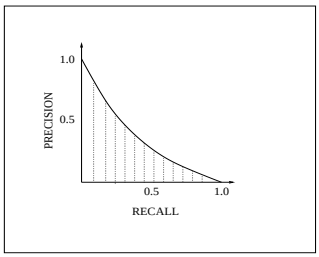
Using the weights (contd.)

- **Given distance of unseen instance d**
  - Binary search on Δ_i array
  - For support vectors with d less than Δ_i, $\gamma_{i_{upper}}$ used, $\gamma_{i_{lower}}$ for others
- **Sum associated with cluster center not computed per-instance**
- **Pre-computed at every position of sorted array**

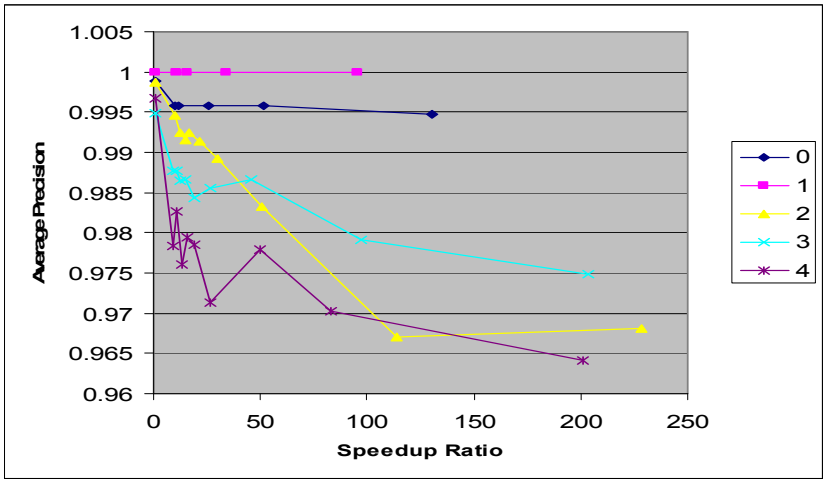


Experiments

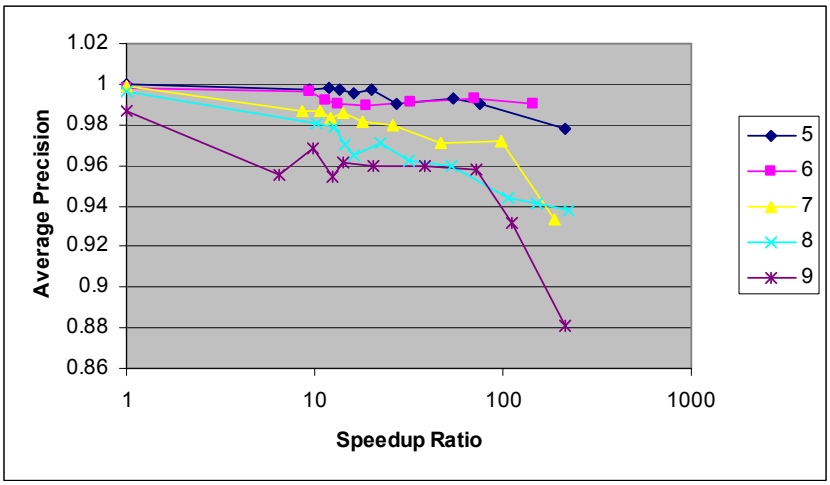
- Datasets
  - TREC video datasets (2003 and 2005)
    - 576 features per instance
    - > 20000 test instances overall
  - MNist handwritten digit dataset (RBF kernel)
    - 576 features
    - 60000 training instances, 10000 test instances
- Performance metrics
  - Speedup achieved over evaluation with all support vectors
  - Average precision achieved



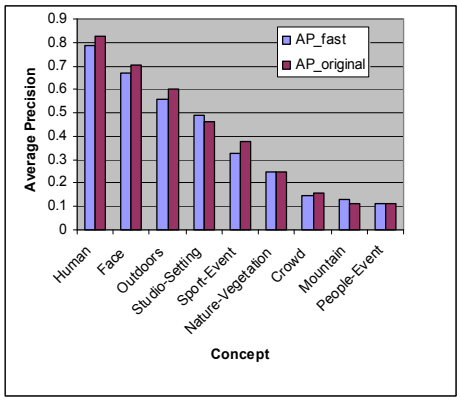
Results (Mnist 0-4)

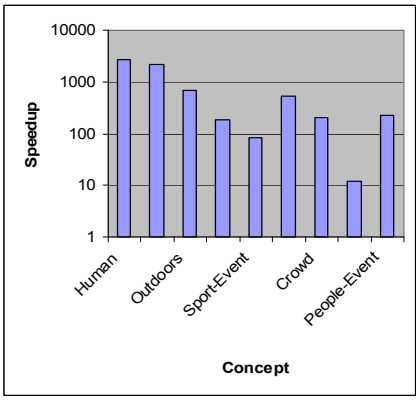


Results (Mnist 5-9)

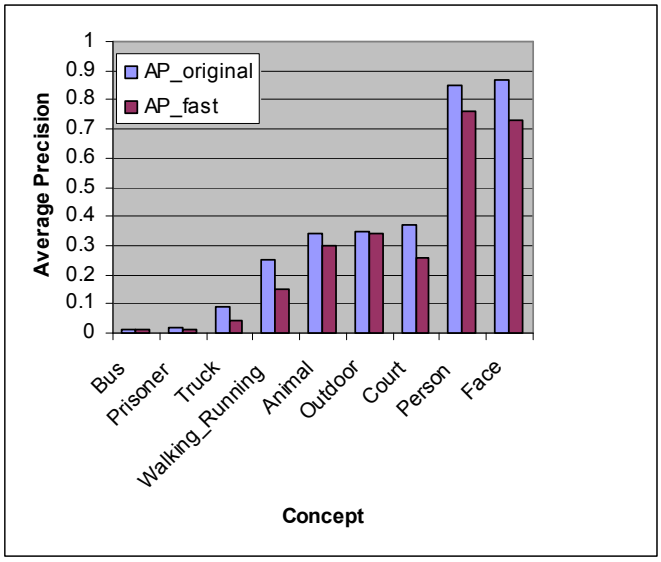


Results (TREC 2003)





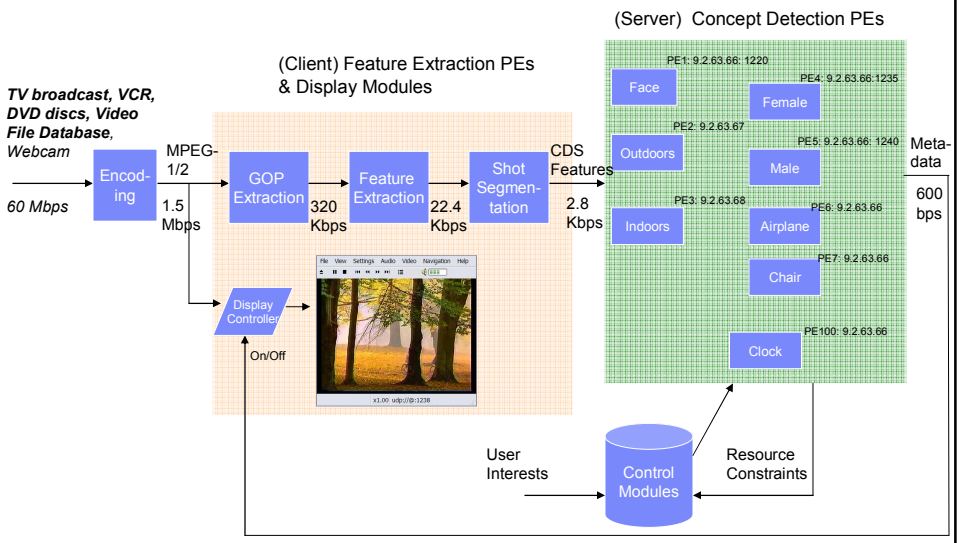
TREC 2005



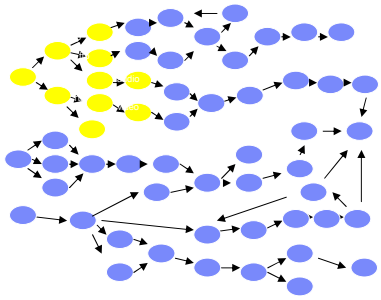
Summary of Complexity Reduction

- Techniques presented demonstrate reasonable performance in terms of both speedup and average precision over multiple concepts in datasets
- Speedups
  - MNist : All concepts at least 50 times faster with AP within 0.04 of original
  - TREC 2003: Eight out of nine concepts speedup greater than 80 times with AP within 0.05 of original
  - TREC 2005: APs in some cases along with speedup respectable
- APs of most concepts close to original APs

VideoDIG Demo
– configurable and scalable video semantic concept detection/filtering



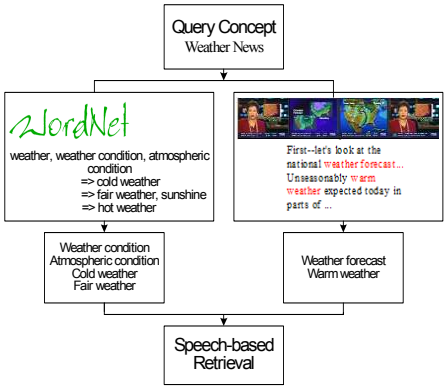
Outline -- Large-Scale Video Semantic Filtering



- Introduction and Motivation
- Prior Arts on Video Semantic Classification
- Compressed-Domain Feature Extraction
- Complexity-Accuracy Trade-Off
- Speed Up SVM Classification
- Speech-Text Complexity-Accuracy Analysis
- Conclusion
- Prospective Projects

Other Issues: Speech and Text-based Topic Detection

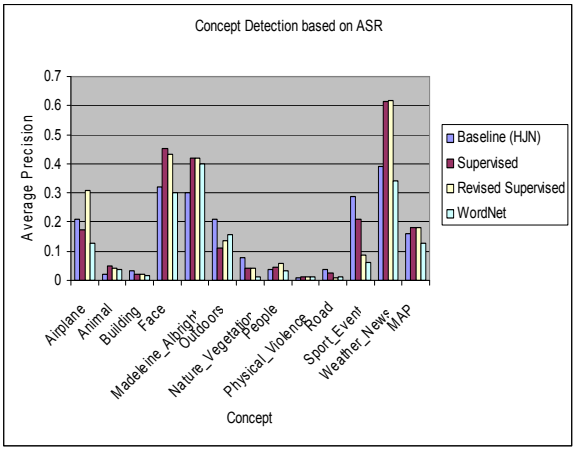
- **Unsupervised learning from WordNet:**



Airplane	Animal	Building	Weather news
airplane	animal	building	Weather
aeroplane	beast	edifice	atmospheric
plane	brute	walk-up	cold weather
airline	critter	butchery	freeze
airbus	darter	apart build	frost
twin-aisle	peeper	tenement	sunshine
airplane	creature	architecture	temper
amphibian	Fauna	call center	scorcher
biplane	microorganism	sanctuary	sultry
passenger	poikilotherm	Bathhouse	thaw
fighter			warm
aircraft			downfall
			hail
			rain

Example of Topic – Related Keywords

Concept Detection based on Automatic Speech Recognition



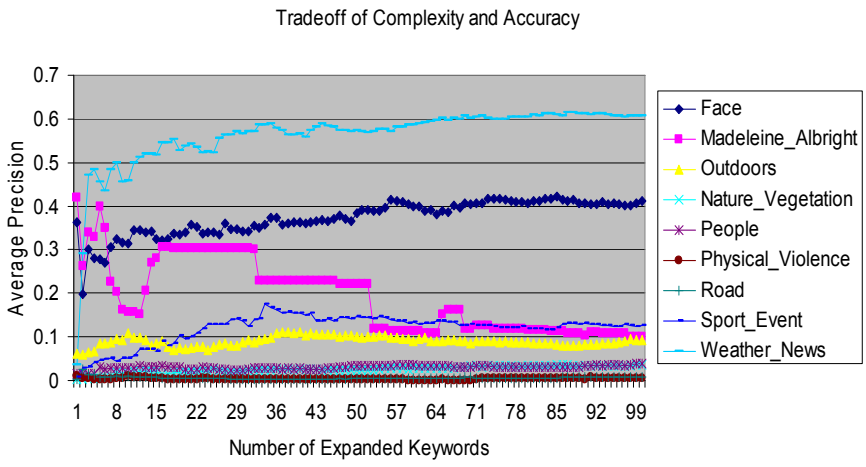
Supervised:

rain 334, temperatur 327, continu 287, weather 284, forecast 269, coast 254, shower 254, storm 245, california 242, northern 226, thunderstorm 225, snow 224, west 208, move 206, lake 195, southeast 187, northeast 175, northwest 172, heavi 170, eastern 161, plain 158, southern 157, expect 153, rocki 153, high 151, warm 147, gulf 143, headlin 142, look 142, great 140, caribbean 140, todai 139, seventi 137, east 135, eighti 134, thirti 133, hawaii 129, meteorologist 125, midwest 125

WordNet:

weather, weather condition, atmospheric condition, cold weather, freeze, frost, fair weather, sunshine, temperateness, hot weather, scorcher, sultriness, thaw, thawing, warming, precipitation, downfall, fine spray, hail, rain, rainfall, monsoon, rainstorm, line storm, equinoctial storm, thundershower, downpour, cloudburst, …..

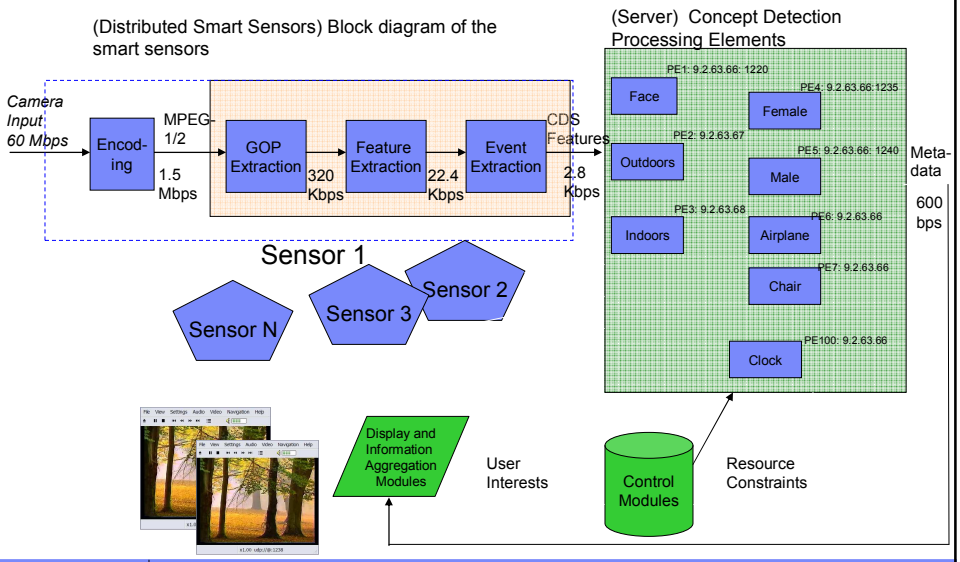
Complexity-Accuracy on Concept Detection based on Automatic Speech Recognition



Conclusion

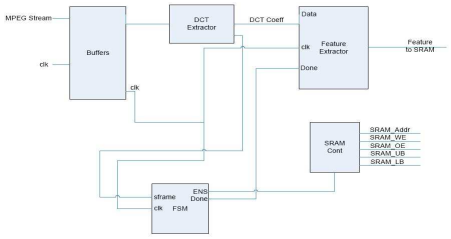
- **Novel Challenge:**
- **Current Status:**
  - Classifying 40 video streams of 100 concepts in real-time by using a Pentium III PC.
  - SVM classification can be speed up in a factor of 10s to 1000s with average precision of 90% or above of the original.
- **Future Works:**
  - Learning and Classifying in noisy environment
  - Frameworks for other types of media (e.g., speech, email, web access, ….) and other types of machine learning/pattern recognition framework (e.g., dynamic Bayesian Net..)

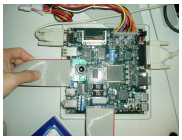
Revisited Video Semantic Filtering Framework



Projects Available (Spring 2006)

- **Developing Distributed Smart Video Cameras (Phase II):**





  - **Large-Scale Data Mining, Community Modeling, Human Behavior Modeling**
  - *Also Course: EE6886 Multimedia Security Systems*

Acknowledgements

- Navneet Panda (UCSB), Xiaodan Song (UW), Victor Sutan (CU), Jason Cardillo (CU)
- Olivier Verscheure (IBM), Lisa Amini (IBM)

Questions?

Contact: cylin@ee.columbia.edu
http://www.research.ibm.com/people/c/cylin